Validation of an STR peak area model

Robert G. Cowell

Faculty of Actuarial Science and Insurance

Sir John Cass Business School

City University London

106 Buhnill Row, London EC1Y 8TZ, UK.

Tel +44 (0)20 7040 8454

Fax +44 (0)20 7040 8572

email rgc@city.ac.uk

January 6, 2009

Abstract

In analyzing a DNA mixture sample, the measured peak areas of alleles of STR markers amplified using the polymerase chain-reaction (PCR) technique provide valuable information concerning the relative amounts of DNA originating from each contributor to the mixture. This information can be exploited for the purpose of trying to predict the genetic profiles of those contributors whose genetic profiles are not known. The task is non-trivial, in part due to the need to take into account the stochastic nature of peak area values. Various methods have been proposed suggesting ways in which this may be done. One recent suggestion is a probabilistic expert system model that uses gamma distributions to model the size and stochastic variation in peak area values. In this paper we carry out a statistical analysis of the gamma distribution assumption, testing the assumption against synthetic peak area values computer generated using an independent model that simulates the PCR amplification process. Our analysis shows the gamma assumption works very well when allelic dropout is not present, but performs less and less well as dropout becomes more and more of an issue, such as occurs, for example, in Low Copy Template amplifications.

Keywords

PCR-amplification; peak area; gamma model.

1 Introduction

Analysing mixed DNA STR profile samples is acknowledged to be a complex task [3]. The importance of using peak-area information (or, alternatively, peak height information: peak area values are known to be highly correlated with peak height values) to analyse mixed DNA STR profile samples is recognised. In recent years, a variety of methods has

been proposed to exploit peak area information, see for example [4] [5] [6]. This paper examines the probabilistic model developed and applied by Cowell, Lauritzen and Mortera in a series of recent papers [1], [7], [8]. In particular, we focus on checking the assumption of using Gamma distributions to model the probabilistic variation in peak area values obtained from the PCR process. Ideally, this would be carried out using data obtained from an extensive controlled experiment involving amplifications of many DNA samples. However, this is not practical. Instead we opt for a simulation approach, in which we simulate peak area values using the stochastic model in [2].

The plan of the paper is as follows. The next section summarises the mathematical model presented in [1] and the simulation model of [2], but ignoring the complication of stutter. Then we examine the fit of the gamma model to simulated data, and also make some analytic comparisons. We then summarise our results. A set of Appendices containing technical mathematical proofs is available as an online supplement.

2 Models and methodology

In this section, we present a brief summary of the gamma model of [1], a summary of the PCR simulation model of [2], and a discussion of the application of the simulation model to testing the gamma model.

2.1 Summary of peak area model for mixtures

The gamma model of [1] considers I potential contributors to a DNA mixture. Let there be M markers to be used in the analysis of the mixture, and that marker m has A_m allelic types, $m = 1, \ldots, M$. Let θ_i denote the proportion of DNA from individual i *prior* to PCR amplification, with $\theta = (\theta_1, \theta_2, \ldots, \theta_I)$ denoting the vector of proportions from all contributors. It is assumed that this pre-amplification proportion of DNA is constant

across markers.

For a specific marker m, the *peak weight* W_{+a} is the *peak area* at allele a multiplied by the *allele number*. This is a simple correction for preferential amplification of alleles. The model is idealized in that it ignores complicating artifacts such as stutter, drop-out alleles and so on, and makes the following further assumptions:

- If W_{ia} denotes the contribution of individual i to peak weight at allele a, then

$$W_{+a} = \sum_i W_{ia}.$$

- W_{+a} is approximately proportional to the amount of DNA of type a.

The key modelling assumption is that each contribution W_{ia} from individual i to peak weight at allele a, is assumed to have a *Gamma distribution*:

$$W_{ia} \sim \Gamma(\rho\gamma_i n_{ia}, \eta)$$

where

- $\gamma_i = \gamma\theta_i$ is the amount of DNA from individual i in mixture;
- n_{ia} is the number of alleles of type a carried by individual i;
- η determines scale and ρ is the amplification factor. Both may be marker dependent.

It follows from the gamma distribution assumption that

$$W_{+a} = \sum_i W_{ia} \sim \Gamma(\rho \sum_i \gamma_i n_{ia}, \eta)$$

and

$$W_{++} = \sum_a W_{+a} = \sum_a \sum_i W_{ia} \sim \Gamma(2\rho\gamma, \eta).$$

By scaling the weight of each allele by the total for the marker we obtain *relative weights* $R_a = W_{+a}/W_{++}$, where it follows that

$$R_a \sim Dir(\rho B_a),$$

where $B_a = \sum_i \gamma_i n_{ia}$ is the weighted allele number, and $B_+ = 2\gamma$ is twice the total amount of DNA γ and is marker independent. Note that

$$\begin{aligned} \mathbf{E}(R_a) &= \mu_a = B_a/B_+ = \sum_i \theta_i n_{ia}/2, \\ \mathbf{V}(R_a) &= \mu_a(1-\mu_a)/(2\rho\gamma+1) = \sigma^2\mu_a(1-\mu_a). \end{aligned}$$

where we define $\sigma^2 = 1/(2\rho\gamma+1)$. Using these moments, a moment-matching approximation to the gamma model using normal distributions can also be developed; this is described in [7].

The attractiveness of the gamma model, and its normal approximation, is that it provides a framework in which a variety of complex mixture problems can be both formulated and solved using Bayesian network technology.

2.2 Summary of simulation model (excluding stutter)

Building on the binomial model of [9], Gill et.al [2] developed a simulation model of the process of analysing STRs. In this section we present a simplified version of their model that excludes stutter. (In Appendix C we present their extended model with a correction for dealing with stutter, and also including dropout occurring through silent alleles.)

The simulation model of PCR amplification without stutter is much simpler than the one with stutter, because it is possible to treat each allele independently of other alleles. Thus, let us concentrate on a sample containing N alleles of type a bound within cells.

The simulation model then has the following three steps that are readily implemented in a small computer program:

Step 1 Extract the alleles from the cells: It is assumed that each allele a has a probability $\pi_{extraction}$ of surviving this process, independently of other alleles, hence the total number of type a alleles n has a binomial distribution:

$$n \,|\, N \sim Binom(N, \pi_{extraction}).$$

Step 2 A portion $\pi_{aliquot}$ of the extracted sample is used for PCR amplification. It is assumed that an individual allele that has survived the extraction process will be randomly selected with probability $\pi_{aliquot}$ for amplification, independently of other alleles, hence the total number of available for amplification, n_0, has a binomial distribution:

$$n_0 \,|\, n \sim Binom(n, \pi_{aliquot}).$$

Note that this may be combined with the distribution in Step 1, to give

$$n_0 \,|\, N \sim Binom(N, \pi_{extraction}\pi_{aliquot}).$$

Step 3 The surviving alleles are then subject to K cycles of amplification. During each amplification cycle, each allele is either duplicated with probability π_{PCReff} or not duplicated with probability $1 - \pi_{PCReff}$. Assuming that alleles do not interfere with each other in the amplification process the number of alleles after the kth cycle, and that the replication probability π_{PCReff} is the same for each cycle, then n_k is related to the number n_{k-1} in the previous cycle by the following recurrence relation:

$$n_k = n_{k-1} + Binom(n_{k-1}, \pi_{PCReff}).$$

For standard amplifications, K is set to 28 cycles, whilst for low copy number procedures, appropriate with $N < 20$, $K = 34$ simulated cycles are used. Using experimental data based on their laboratory procedures, [2] estimated the following values for the various probability parameters $\pi_{extraction} = 0.6$, $\pi_{aliquot} = 20/66$ and $\pi_{PCReff} = 0.82$. We use these values in all our simulations. Finally, the alleles register a peak if their number exceeds the threshold total of 2.35×10^7, and numbers are scaled by the factor of 2×10^6 to convert them to a representative peak area value.

Note that the iteration in Step 3 defines a Galton-Watson branching process. Such processes have been used to model aspects of PCR by other authors, for example [10] who studied mutations in the PCR process, and [11] who estimate the efficiency of the amplification process.

2.3 Nature of simulations

There are some differences in the simulation model and the gamma model that need to be accommodated in order to use the simulated peak areas to test the gamma model. The gamma model introduces weighted areas, that is, the weight associated with a peak area of an allele having repeat number a is the peak area times the repeat number. This is to accommodate preferential amplification, where longer alleles tend to amplify less efficiently than shorter alleles. Such preferential amplification is not modelled in the simulation model, as the amplification efficiency parameter π_{PCReff} does not depend on allele length. Therefor, in our comparison we shall model the peak areas directly using gamma distributions, rather than the weighted peak areas: that is, interpret the W_{ia} etc. in Section 2.1 directly as peak areas and not peak weights. Secondly, the gamma model does not deal with thresholds, whereby in PCR analysis of real DNA samples a peak height has to be above a certain threshold to be counted as a peak. Hence in simulating the peaks areas we shall set the threshold number to zero, and merely scale the counts by 2×10^6 to obtain

an area value.

This calibration factor, converting simulated counts to area values, is not ideal because it does not give the correct range of values for peak areas that one would expect to see in terms of RFU units from an electropherogram. However our analysis will not depend on the precise value that the calibration should take for two reasons. The first is that a positive multiple of a gamma distributed random variable is also a gamma distributed random variable. Hence in testing whether the gamma distribution family is suitable to use to model peak areas, the value of the calibration factor is not an issue. The second reason why the value of the calibration factor is not important for this paper is that the likelihoods that are derived from the peak area values that enter the probabilistic model depend on the relative areas R_a, and in forming these ratios of areas, the scaling calibration cancels out.

Finally the gamma model does not deal with allelic dropout. In the simulation model, dropout of allele a occurs either because no alleles of type a survive the first two stages of the simulation–extraction and subsequent sampling–or if some do they fail to amplify beyond the threshold. As we are setting the threshold to zero, this means that dropout can only happen in our simulations if an allele of a certain type a fails to survive the first two steps of the simulation. Therefor in simulating many samples, we shall only retain those in which dropout does not occur. We shall return to these issues in more detail in Section 4.

It is worthwhile emphasizing here the distinct natures of the gamma model and the simulation model. The simulation model of [2] generates mixture peak areas that appear to capture the important statistical characteristics of peak area values obtained from PCR amplification of real DNA mixture samples, incoporating artifacts such as stutter and dropout. The simulation method itself is not used to analyse any given mixture, for example to identify unknown genetic profiles of individual contributors. In contrast, the

gamma model has been developed for the purpose of analysing given mixtures using the peak area information. Calculations based on the gamma model can be carried out quickly and accurately using probabilistic expert system software. The gamma model presented and analysed in this paper does not cater for stutter or dropout, so where such effects start to become dominant in the simulated samples we could expect the gamma model to start breaking down, and indeed we shall see that this is the case. Work is in progress to extend the gamma model to cater for artifacts, and it is hoped to present this extension elsewhere. The aim of this paper is to see if the gamma model provides a good fit to situations in which artifacts are not present: if it does then it makes sense to try and extend the model to cope with artifacts. On the other hand if the gamma model turns out to fit poorly in the simplified situation of no artifacts, then one would not expect an extended gamma model to fit any better. As we shall see, the gamma model fits well when artifacts are not present, justifying the attempt to extend it to cater for artifacts.

3 Results

Given the restrictions described above, we proceed to use the simulation model to test the following aspects of the gamma model:

Peak areas The peak areas follow gamma distributions with mean proportional to the amount of DNA in the sample, and scale parameter independent of the amount of DNA in the sample.

Relative areas The gamma model predicts that the relative areas $R_a = W_{+a}/W_{++}$ follow Dirichlet distributions $R_a \sim Dir(\rho B_a)$, where $B_a = \gamma \sum_i \theta_i n_{ia}$. For the special case of amplifying the DNA of a single heterozygote person, the prediction simplifies to $R_a \sim Beta(2\rho\gamma)$ having mean $\mathbf{E}(R_a) = 0.5$ (because no allowance has been made for preferential amplification) and variance $\mathbf{V}(R_a) = 1/4(2\rho\gamma+1)$. We examine how well

the Beta distribution fits, and also examine the behaviour of the variance: according to the model the *inverse* of the variance will be linearly proportional to the amount γ of DNA prior to amplification.

We shall look at each aspect in turn.

3.1 Peak area size

In Figure 1 is plotted the results of simulating counts (representing areas) for three different starting values of N, the number of alleles so a certain type. For each value of N, 10,000 amplifications were simulated, and those that did not lead to dropout were retained. The histogram plots in the first column indicating more skewness in the distribution the lower the value of N. All indicate that gamma distributions are appropriate. This is confirmed in the second column of plots, in which moment estimation was used to estimate the parameters of the gamma distribution for each N, and values simulated from this to construct the quantile-quantile plots in the second column. All plots are highly linear over the range except for departures from linearity apparent to the top right hand corners of the plots. In [7] a normal approximation to the gamma model was suggested. Hence in the third column is plotted normal quantile-quantile plots based on the simulated counts. All plots are reasonable linear over the range $(-2, 2)$, but departures are apparent outside this range for the two simulations with starting values $N = 5$ and $N = 20$, consistent with the skewed histograms. It is worth pointing out that the plots in Figure 1 were based on 28 simulated PCR cycles, but starting values $N = 5$ and $N = 20$ correspond to low copy number situations, for which 34 simulated cycles should be used [2]. However, using 34 simulated cycles leads to similar plots (not shown) to those in Figure 1.

FIGURE 1 HERE

A second assumption of the gamma model is that the mean area associated with an allele is proportional to the amount of DNA of that allelic type in the sample prior to amplification : $W_{ia} \sim \Gamma(\rho\gamma_i n_{ia}, \eta)$. This assumption is verified by the plot in Figure 2, (10000 simulations for each integer value of $N \in (1, 100)$) which shows a strong linear dependence of mean area on the amount over a large range, with just a hint of departure from linearity in the low copy number regime. This result can be derived analytically for the simulation model, as shown in Appendix A.

FIGURE 2 HERE

Another assumption of the gamma model is that, for each marker, the scale parameter η does not depend on the amount. To test this, scale parameters were estimated for each value of N using the same simulations as for Figure 2. The plot in Figure 3 shows that the common scale parameter assumption appears reasonably valid for the range $N > 25$, but the assumption appears to breakdown in the range of values of N in the low copy number regime. The breakdown is arising in the variation arising from the selection of alleles in the pre-PCR stage. For the PCR stage itself the assumption of constant scale parameter η holds as follows. It is shown in Appendix A that the mean number of alleles arising from r simulated PCR cycles starting from a single allele is $E_r[X] = (1 + \pi_{PCReff})^r$, and the variance is $V_r[X] = (1 - \pi_{PCReff})(1 + \pi_{PCReff})^{r-1}((1 + \pi_{PCReff})^r - 1)$. For an initial number N alleles amplifying independently, both the mean and variance are multiplied by N. (This follows because of the independence. Alternatively, it can be found by finding the moments using $F_r(t)^N$.) For the gamma distribution, the moment matching estimate of the scale parameter η is $E_r[X]/V_r[X] = (1-\pi_{PCReff})((1+\pi_{PCReff})^r - 1)/(1+\pi_{PCReff})$ which does not depend on N.

FIGURE 3 HERE

3.2 Relative area size

We have seen that the use of gamma distributions appears to be justified from the simulations in Section 3.1. We now look at the gamma model prediction that the relative peak areas obtained by amplifying the alleles of a heterozygote individual are Beta distributed. In Figure 4 we plot the results of simulations of relative areas for three different starting values for the number of alleles of each type. Each simulation sampled 10000 amplification values, only those that did not result in a dropout were retained for calculating relative area values. To the left are histograms, indicating a symmetric distribution with mean 0.5. On the right are corresponding quantile-quantile plots which use the simulated values for actual quantiles, plotted against quantiles of values simulated from a beta distribution of mean 0.5 and variance matching that of the relative area values. All quantile-quantile plots yield close to straight lines, indicating that the beta distribution is a very good fit to the data, even for the low-copy number regime with a starting value of 5 for each type of allele.

FIGURE 4 HERE

Figure 5 shows the results of a similar simulation, in which the starting number of alleles was unequal, in a ratio $n_a : n_b = 1 : 3$. This could correspond to two homo-zygote contributors contributing DNA in a ratio of 1:3 to a mixture, or one homo-zygote and one heterozygote having one allele in common and contributing equal amounts to a DNA mixture. Again, the quantile-quantile plots fit straight lines well, indicating that the relative areas are following appropriate Beta distributions.

FIGURE 5 HERE

A second prediction of the gamma model regarding the relative areas is that the inverse of the variance of the relative area is linearly proportional to the starting amount of DNA.

From the histogram plots in Figure 4 the variance does appear to decrease as n_a increases. For a fuller analysis of the dependence of variance on amount, one hundred amplifications were simulated of a heterozygote, with each of the starting values $n_a(= n_b) = 1, 2, \ldots, 100$ for the number of alleles to each type, with 28 simulated PCR cycles. From each simulation the variance of the relative areas was calculated for those simulated amplifications that did not lead to dropout. The variance and inverse variance are plotted against n_a (which is proportional to the amount of DNA in the gamma model) in Figure 6.

FIGURE 6 HERE

From the plot of inverse variance, a good straight line fit is apparent for the range $n_a \geq 20$ (the sample correlation coefficient of the values in this range is 0.998). Thus, the gamma model is providing an excellent fit over this range. However it is clear than for smaller amounts the gamma model prediction regarding the dependence of variance on amount is breaking down. The plot on the right reaches a minimum for $n_a = 14$, and for lower values of n_a the inverse variance is increasing, corresponding to the variance decreasing with decreasing n_a for $n_a \leq 14$.

This somewhat surprising inflexion in the curve can be understood by a closer examination of how variability arises from the simulation model. Recall there are two stages to the simulation process, the pre-PCR selection stage (consisting of two steps) and the PCR stage. Each stage provides variability. In the pre-PCR selection stage, alleles are sampled for the PCR stage, each with probability $p = 0.6/3.3 = 0.182$. Thus an equal number $n_a = n_b$ of alleles of types a and b can easily become unbalanced prior to the PCR stage of the simulation, and the PCR amplification stage can further exacerbate the in-balance.

Now consider the extreme case where we start out with $n_a = n_b = 1$. The simulations only retain those values where there is no dropout. This requires that both alleles are present for the simulation of the PCR stage, and thus start out with the same number,

that is $n_a = n_b = 1$, and the ratio of one allele count to the total is 0.5. Thus there is no pre-PCR variability in the relative amounts of the two alleles entering the amplification stage, and so the total variability is less than may be expected at first sight. On the other hand, for very large values of $n_a = n_b$, the Beta distribution become very highly peaked and narrow, so that the values sampled from $Binom(n_a, p)$ and $Binom(n_b, p)$ are close to $n_a p$ and $n_b p$, leading to a relative values close to 0.5 and again a small variance that goes to zero as $n_a = n_b \to \infty$. For the intermediate values of $n_a = n_b$ both sources of variability are present, and so by continuity we should expect a maximum value somewhere. For the parameters used in the simulation the greatest variance occurs at $n_a = 14$, as shown in Appendix B for the pre PCR sampling stage, and also $n_a = 14$ in Figure 6 where both sources of variation are combined.

4 Discussion

This paper has presented the results of a simulation study of the gamma model of [1] for the distribution of peak areas arising in the PCR amplification of STRs. The study has been based on a simplified version of the model of [2] for simulating the whole process of amplifying STRs to create realistic simulations of peak area distributions. On the whole, the gamma model is seen to compare favourably for modelling the peak areas simulated from the model of [2]. The plots in Figure 1 show that gamma distributions model the simulated areas well, even in the low copy number regime ($N \leq 20$). The normal approximation suggested in [7] works well for large N but breaks down in the low copy number regime. We have shown in Appendix A that the assumption of the linear dependence of the mean of the gamma distribution on the amount is a consequence of the simulation model. There is a slight departure for the low copy number regime that can be explained by the pre PCR sampling of alleles in the extraction and aliquot

selection processes that can lead to some or all alleles of specific type not being selected for amplification. The gamma model does not cater for such dropout behaviour. The assumption of the common scale parameter η has been shown to hold via simulation for large N values but appears to break down in the low copy number regime. Again this can be traced to the pre PCR sampling of alleles for the amplification process. The results of Section 3.2 shows that the relative areas of two alleles in a simulated amplifications closely follow beta distributions even in the low copy number regime. The prediction of the gamma model that the variance of the relative areas should increase as the amount of DNA decreases holds in the regime for large N, but breaks down in the low copy number regime. The explanation of this behaviour in the simulation model is given in detail in Appendix B, and again lies in the variability of alleles introduced by the pre PCR sampling of alleles for amplification.

We must now return to the points made in Section 2.3, and address the relevance of the simulation study of this paper to the application of the gamma model to the analysis of real STR mixtures.

The main issue is that the gamma model of [1] used a simple correction for preferential amplification that the present simulation study did not use, because the simulation model of [2] does not include simulation of preferential amplification. The simple correction of [1] does not take into account the dependence of preferential amplification of DNA on the condition of the sample. The question arises as to whether it provides a good enough approximation: ultimately this can only be resolved by experimental analysis of real STR PCR amplifications.

Another issue concerns stutter. The simulations reported in this paper used a simplified model that ignored the generation of stutter peaks. This was because the model of [1] does not model stuttering. However an extension to take account of stuttering is being developed and will be reported elsewhere. Another property of the simulations used this paper is that

they were conditional on no dropout occurring. This was appropriate for testing the model of [1] as it does not take account of dropout. Real PCR amplifications do lead to dropout, and there appear to be three main causes. These are:

1. Pre-selection of alleles lead to some alleles types not being selected for the PCR amplification stage. This is particularly acute for low copy number amplification.

2. Silent alleles: alleles that do not amplify because the primer does not bind to the flanking regions of thee DNA.

3. Alleles are amplified, but below a threshold for detection.

Of these possibilities, extensions of the gamma model to incorporate dropout due to (1) and (2) are being developed and will be reported elsewhere. The problem of modelling the detection threshold has still to be addressed withing the context of the (extended) gamma model.

In summary, the gamma model provides a good approximate fit to data generated using the simulation model when the initial number of alleles N is relatively large so that the stochastic variation introduced by the selection steps 1 and 2 described in Section 2.2 are much smaller that arising from the amplification process of step 3. In contrast the fit of the gamma model breaks down in the low copy number regime, and the lack of fit can be traced to the pre-PCR selection stage of the simulation model which is not captured by the gamma model. We agree with Gill et al. [2] that this pre-PCR selection stage can be a dominant source of peak imbalance and dropout in real DNA samples when the initial number of alleles in a DNA sample is initially small. However, extensions of the gamma model to incorporate dropout to take account of this are being developed. The present study has shown that the simple gamma model of [1] analyzed in this paper forms a solid foundation upon which to build such extensions.

Acknowledgements

The author would like to thank Professors S. L. Lauritzen and J. Mortera for helpful discussions and encouragement regarding this work, and two referees whose suggestions have improved the paper.

References

[1] Robert G. Cowell, Steffen. L. Lauritzen, and Julia Mortera. A gamma model for DNA mixture analyses. *Bayesian Analysis*, 2:333–348, 2007.

[2] Peter Gill, James Curran, and Keith Elliot. A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. *Nucleic Acids Research*, 33(2):632–643, 2005.

[3] P. Gill, C. Brenner, J. Buckleton, A. Carracedo, M. Krawczak, W. Mayr, N. Morling, M. Prinz, P. Schneider, and B. Weir. DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International*, 160:90–101, 2006.

[4] Ian W. Evett, Peter D. Gill, and James A. Lambert. Taking account of peak areas when interpreting mixed DNA profiles. *Journal of Forensic Sciences*, 43:62–69, 1998.

[5] M. W. Perlin and B. Szabady. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *Journal of Forensic Sciences*, 46:1372–1378, 2001.

[6] M. Bill, P. Gill, J. Curran, T. Clayton, R. Pinchin, M. Healy, and J. Buckleton. PENDULUM — a guideline - based approach to the interpretation of STR mixtures. *Forensic Science International*, 148:181–189, 2005.

[7] Robert G. Cowell, Steffen. L. Lauritzen, and Julia Mortera. Identification and separation of DNA mixtures using peak area information. *Forensic Science International*, 166:28–34, 2007.

[8] Robert G. Cowell, Steffen. L. Lauritzen, and Julia Mortera. Probabilistic modelling for DNA mixture analysis. *Forensic Science International: Genetics Supplement*, to appear. Available electronically from 23 April 2008.

[9] Gustavo Stolovitzky and Guillermo Cecchi. Efficiency of DNA replication in the polymerase chain reaction. *Proceedings of the National Academy of Science, USA*, 93:12947–12952, 1996.

[10] Fengzhu Sun. The polymerase chain reaction and branching processes. *Journal of Computational Biology*, 2:63–85, 1995.

[11] Nadia Lalam, Christine Jacob, and Peter Jagers. Estimation of the PCR efficiency based on a size-dependent modelling of the amplification process. *C. R. Acad. Sci. Paris, Ser I*, 341:631–634, 2005.

[12] Peter Gill, James Curran, and Keith Elliot. Supplementary material. Published online at Nucleic Acids Research, 2005.

Figure captions:

Figure 1 Simulations of area amplifications for three different starting values N of the number of an alleles, using 28 amplification cycles.

Figure 2 Mean area as a function of starting number N of alleles, using 28 amplification cycles.

Figure 3 Estimated scale parameter η as a function of amount N of starting number of alleles.

Figure 4 Simulations of relative areas for amplification of hetero-zygotic individual, for three different starting values of the number of alleles of each type

Figure 5 Simulations of relative areas for amplification for a 1:3 imbalance of number of alleles of either type to start with.

Figure 6 Variance and inverse variance of relative area R_a plotted against starting number of alleles n_a for the amplification of a hetero-zygotic individual.

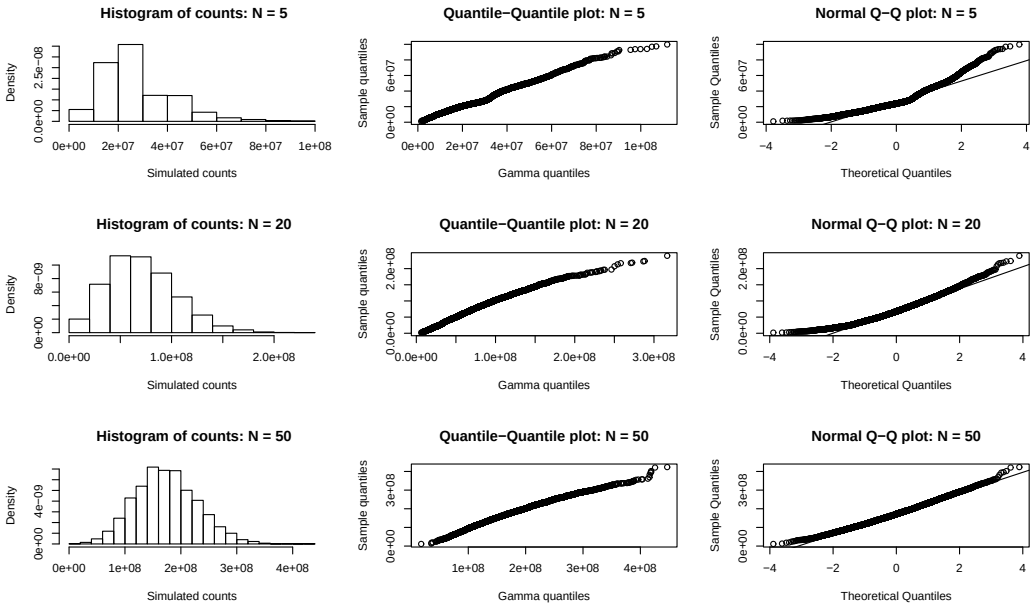

Figure 1: Simulations of area amplifications for three different starting values N of the number of an alleles, using 28 amplification cycles.

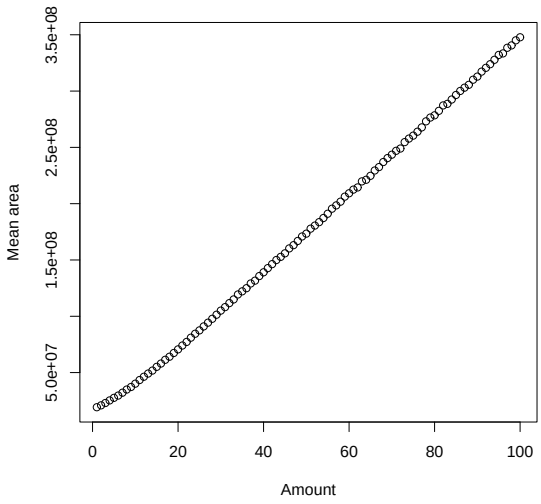

Figure 2: Mean area as a function of starting number N of alleles, using 28 amplification cycles.

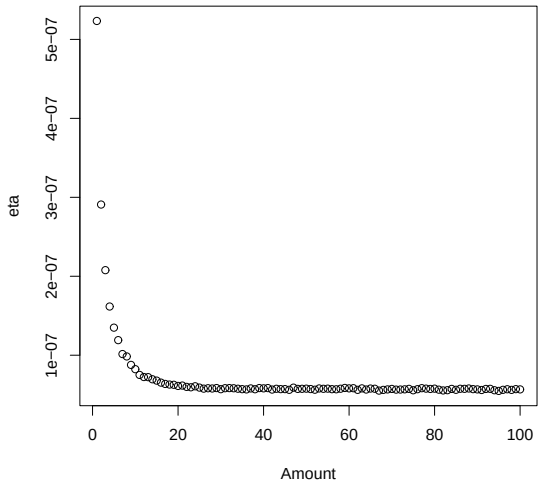

Figure 3: Estimated scale parameter η as a function of amount N of starting number of alleles.

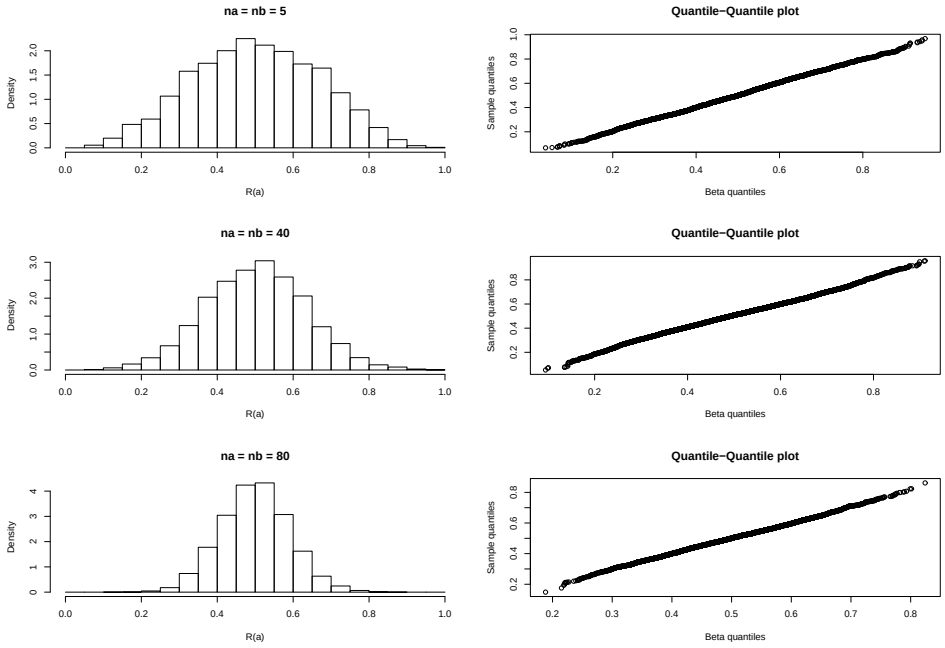

Figure 4: Simulations of relative areas for amplification of hetero-zygotic individual, for three different starting values of the number of alleles of each type.

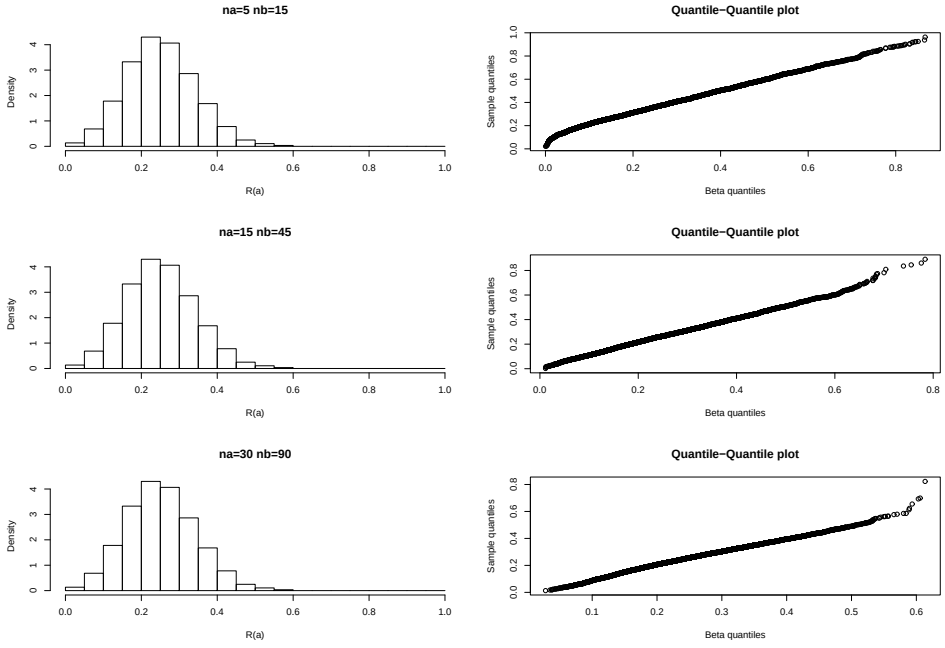

Figure 5: Simulations of relative areas for amplification for a 1:3 imbalance of number of alleles of either type to start with.

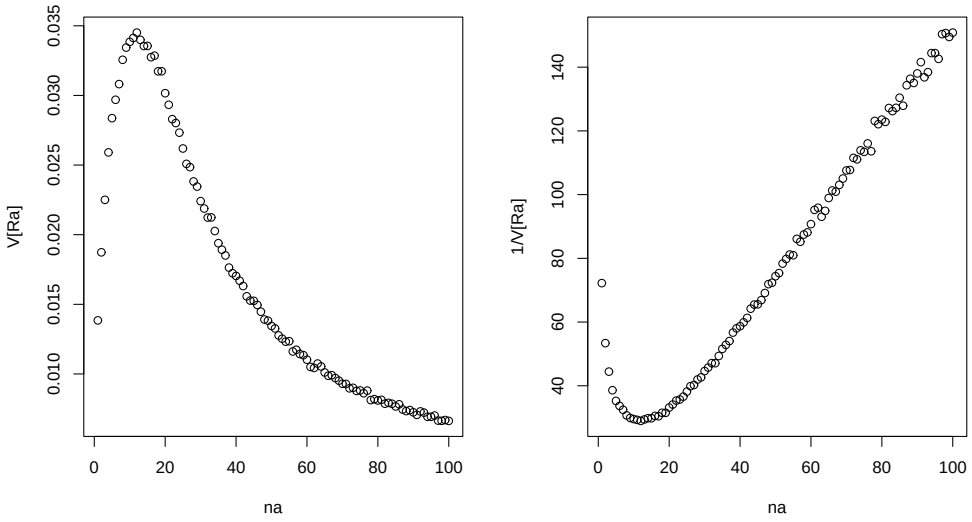

Figure 6: Variance and inverse variance of relative area R_a plotted against starting number of alleles n_a for the amplification of a hetero-zygotic individual

SUPPLEMENTARY MATERIAL

A Linear dependence of mean area on amount

In this appendix we derive the probability generating function (PGF) for the distribution of the number of alleles in the simplified simulation model given the starting number of alleles of a given type, and use the PGF to show the linear dependence of the mean on the amount. We follow the approach of [1] based on cascade theory.

Consider a single allelic molecule a, and all of the copies produced from it in the PCR simulation. Call the time before the first PCR cycle the zero-th generation. Let us denote by x_r the number of alleles of type a after exactly r PCR cycles, with $r = 0$ denoting the number of alleles of type a prior to any PCR cycle taking place, so that $x_0 = 1$. Now each individual molecule is either copied or not, independently of the other molecules, so that the PGF for the number of molecules of type a after one amplification cycle will be

$$F(t) = (1 - \pi_{PCReff})t + \pi_{PCReff}t^2. \tag{1}$$

More generally, we have

Theorem A.1 *The PGF for the number of molecules after r PCR cycles, given that there is exactly one prior to any amplification cycle, is*

$$F_r(t) = F(F(F(\ldots F(t)\ldots))) \tag{2}$$

where

$$F_1(t) = F(t), F_{s+1}(t) = F(F_s(t)), \quad (s = 1, 2, 3, \ldots). \tag{3}$$

If there are n_0 molecules of type a to begin with then the PGF for the number after r PCR cycles, assuming independent amplification of alleles (which is assumed by the

simulation model) will be $F_r(t)^{n_0}$. However, the number n_0 is not fixed, but is a random variable that depends on the first two selection steps of the pre PCR stage. If there are N alleles to start with, prior to selection, then as summarised in Section 2.2 (main paper) $n_0 \,|\, N \sim Binom(N, \pi_{extraction}\pi_{aliquot})$. This has PGF given by

$$G(t) = (1 - \pi_s + \pi_s t)^N.$$

where we define $\pi_s = \pi_{extraction}\pi_{aliquot}$. Hence the final PGF for the number of alleles of type a, given there are N to begin with, with r simulated PCR cycles for all three Steps of the simulation model described in Section 2.2 (main paper) is given by

$$H(t) = G(F_r(t)) = (1 - \pi_s + \pi_s F_r(t))^N. \tag{4}$$

It is now a simple matter to find the mean of the distribution, by differentiating with respect to t and setting $t = 1$. Now,

$$\frac{dH(t)}{dt} = N\pi_s(1 - \pi_s + \pi_s F_r(t))^{N-1}\frac{dF_r(t)}{dt} \tag{5}$$

Let $E_r[X]$ denote the mean number of alleles of type a after r amplification cycles starting

from one allele. Then

$$\begin{aligned} E_r[X] &= \frac{dF_r(t)}{dt}\Big|_{t=1} \\ &= \frac{dF(F_{r-1}(t))}{dt}\Big|_{t=1} \\ &= \frac{d\left[(1-\pi_{PCReff})F_{r-1}(t)+\pi_{PCReff}F_{r-1}(t)^2\right]}{dt}\Big|_{t=1} \\ &= (1-\pi_{PCReff})\frac{dF_{r-1}(t)}{dt}+2\pi_{PCReff}F_{r-1}(t)\frac{dF_{r-1}(t)}{dt}\Big|_{t=1} \\ &= (1+\pi_{PCReff})E_{r-1}[X]. \end{aligned}$$

The solution of the recurrence relation, using $E_0[x]=1$, is

$$E_r[X] = (1+\pi_{PCReff})^r, \tag{6}$$

from which it follows that

$$\frac{dH(t)}{dt}\Big|_{t=1} = N\pi_s(1+\pi_{PCReff})^r, \tag{7}$$

which is clearly linear in N.

If $V_r[X]$ denotes the variance in number of alleles of type a after r amplification cycles starting from one allele, then it can be shown that

$$V_r[X] = (1-\pi_{PCReff})(1+\pi_{PCReff})^{r-1}((1+\pi_{PCReff})^r-1) \tag{8}$$

The slight curvature in Figure 2 (main paper) arises because that plot was based on samples for which there is not dropout. Conditioning on no dropout means that the original distribution for the pre-PCR sampling steps yields a truncated Binomial distribution with

PGF given by

$$G(t) = \sum_{i=1}^{N} \frac{\binom{N}{i}(1-\pi_s)^{N-i}\pi_s^i t^i}{1-(1-\pi_s)^N}. \tag{9}$$

The full PGF for the simulation process is now $H(t) = G(F_r(t))$, from which the mean is found to be

$$\frac{N\pi_s(1+\pi_{PCReff})^r}{1-(1-\pi_s)^N}$$

This departs from linearity in N for small values of N.

B Derivation of maximum variance in simulation

In the pre-PCR selection stage of the simulation, if N alleles are available to be sampled for amplification, the number sampled is binomially distributed with $\pi_s = \pi_{extraction}\pi_{alequot} = 0.6/3.3 = 0.182$. Suppose that N alleles of type a and N alleles of type b are available to be selected. Denote by N_a and N_b the number of alleles selected. Then

$$\begin{aligned} N_a &\sim Binom(N, \pi_s) \\ N_b &\sim Binom(N, \pi_s) \end{aligned}$$

The sampling is done independently for each type of allele, thus $P(N_a, N_b) = P(N_a)P(N_b)$. We concentrate on no dropout, thus we are interested in the variance of $N_a/(N_a + N_b)$ conditional on both $N_a > 0$ and $N_b > 0$. This requires the distribution $P(N_a, N_b \,|\, N_a > 0, N_b > 0) = P(N_a \,|\, N_a > 0)P(N_b \,|\, N_b > 0)$ where the conditional probabilities are obtained from the truncated binomial, thus

$$P(N_a \,|\, N_a > 0) = \binom{N}{N_a} \frac{\pi_s^{N_a}(1-\pi_s)^{N-N_a}}{1-(1-\pi_s)^N}, \quad N_a = 1, 2, \ldots, N$$

and with a similar expression for N_b. The conditional variance is readily evaluated numerically for various values of N. Figure 1 plots the variance as a function of $N = 1, 2, \ldots, 25$. The plot reaches a maximum at $N = 14$.

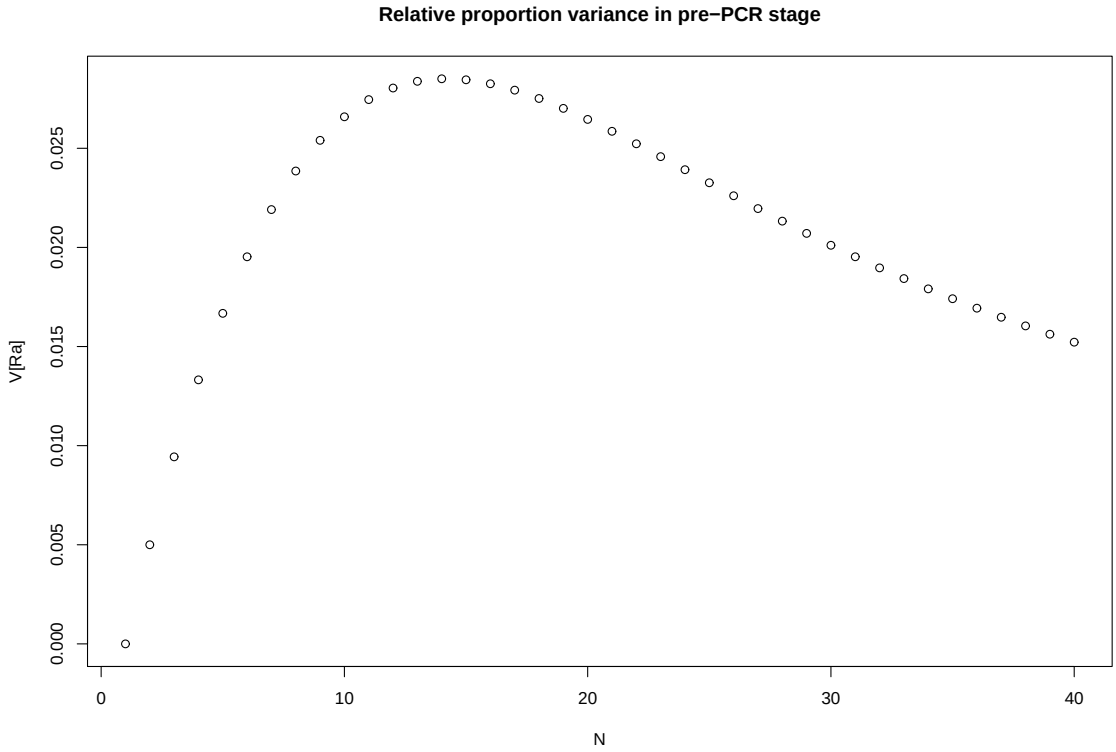

Figure 1: Variance of sampling proportion $N_a/(N_a+N_b)$ in the pre-PCR stage, as a function of the (equal) numbers of each pair of alleles to be sampled.

C Full simulation model, including stutter

The full simulation model of [2] differs from the simplified version used in this paper in that it also models the generation of stutter peaks. Unfortunately their is a small error in their formulation that we correct for here. Their formulae are found in an online supplement [3] [2], and for simplicity we use their notation in this appendix.

Their basic simulation model in Appendix 2 of the supplement consists of the following relations (we only give the relevant equations here). These equations are assuming that there are no alleles of repeat number A to begin with.

1 $n^A_{survived} \sim Bin(N, \pi_{extraction})$

3 $n^A \sim Bin(n^A_{survived}, \pi_{aliquot})$

5 $n_A(k) \sim n_A(k-1) + Bin(n_A(k-1), \pi_{PCReff})$ with $n_A(0) = n_A$

10 $s_A(k) \sim s_A(k-1) + Bin(s_A(k-1), \pi_{PCReff}) + Bin(n_A(k), \pi_{stutter})$, with $s_A(0) = 0$

Together, equations (1), (3) and (5) constitute the simplified simulation model used in this paper. Equation 10 introduces stutter ignored in this paper. s_A denotes the stutter peak size at repeat number $A-1$ arising from the alleles of repeat number A. Equation 10 models the size of the stutter peak after k amplifications is the number at the end of the previous cycle, plus a random number that were present in the previous cycle that get amplified, plus a contribution from a random number of alleles of repeat number A available for the k-th cycle, $n_A(k)$, that mis-copied to yield alleles of repeat number $A-1$. The error in the system of equations is that if an allele of repeat number A mis-copies to a stutter allele $A-1$, it cannot also make a copy of repeat number A. However equation (5) implies that is is available to do so. In essence the Binomial term on equation (5) should be corrected as it overestimates $n_A(k)$. In addition, equation (10) also needs a similar

correction for an allele of repeat number $A-1$ can make a copy of itself, or make a copy having a repeat number of one less, or not make either type of copy. However, using the estimated value $\pi_{stutter} = 0.002$, the difference that these errors make will be slight.

We now describe our full simulation model for the entire process of PCR amplification. We need only treat a single marker, as we simulate amplification of distinct markers independently.

Thus suppose M is a marker system, let A denote the set of alleles in the allelic ladder of M. In this set of alleles there will be at least one allele of repeat number a such that the repeat number $a-1$ is not in A. For example, for TH01, the ladder system could be the set $\{5, 6, 7, 8, 8.3, 9, 9.3, 10, 11, 12, 13, 14\}$. The repeat number 5 is in the set, but 4 is not; similarly 8.3 is in the set, but 7.3 is not. In our simulation model, we assume that such "lower boundary" alleles cannot amplify to make a stutter allele. For our simulation model, we use vectors of counts indexed by the allele set A, one is a running count of alleles $n(a)$ at the various PCR stages, the other $s(a)$ is a temporary vector to take account of stutter. Assume we have I people, and that person i contributes m_i cells to the simulated mixture. Let n_{ia} denote the number of alleles of type $a \in A$ that person i has. Note that if person i has a silent allele then $\sum_a n_{ia} = 0$ if both alleles are silent, and $\sum_a n_{ia} = 1$ if only one allele is silent. If neither is silent then then $\sum_a n_{ia} = 2$. In this way silent alleles a readily incorporated into our simulation model, which we now present in the form of pseudo-code.

Pseudo-code implementation of PCR simulation algorithm:

Initialisation for each $a \in A$ do $\{n(a) := \sum_i m_i n_{ia}$ and $s(a) := 0\}$;

Extraction for each $a \in A$ do $n(a) := Binom(n(a), \pi_{extraction})$;

Aliquot for each $a \in A$ do $n(a) := Binom(n(a), \pi_{aliquot})$;

Amplify

For $k = 1$ step 1 until K do

- for each $a \in A$ such that $a - 1 \in A$ do $s(a-1) := Binom(n(a), \pi_{stutter})$;
- then:
  - for each $a \in A$ do
    - if $a - 1 \in A$ do $n(a) := n(a) + Binom(n(a) - s(a-1), \pi_{PCReff})$
    - otherwise do $n(a) := n(a) + Binom(n(a), \pi_{PCReff})$;
  - for each $a \in A$ do $n(a) := n(a) + s(a)$;
  - for each $a \in A$ do $s(a) := 0$;

In the amplification stage, we first allow alleles to stutter, and put their counts in the vector $s(a)$. We then allow the alleles to make copies of themselves, taking care to avoid the possibility that an allele that has made a stutter copy of itself does not also make an exact copy of itself: the *if* condition deals with this event. We then update the running total in the vector $n(a)$ by adding to it the counts of stutter alleles $s(a)$. We then reset all $s(a)$ to zero ready for the next amplification round. (Notice they were all set to zero in the initialization stage.) The vector $n(a)$ gives the vector of simulated counts of the various alleles at then end of the simulation.

Notice that this simulation model allows stutter products to themselves make stutter products during the amplification cycles. However such sub-stutter counts will tend to be very small in comparison to main peak counts. Also notice that if alleles a and $a - 1$ are both present initially, then the simulation model allows for stutter product $a - 1$ generated by a to be made, but such hidden stuttering will not be obviously apparent as it will be masked by the normal amplification of the $a - 1$ alleles.

References

[1] I. J. Good. The number of individuals in a cascade process. *Proceedings of the Cambridge Philosophical Society*, 45:360–363, 1949.

[2] Peter Gill, James Curran, and Keith Elliot. A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. *Nucleic Acids Research*, 33(2):632–643, 2005.

[3] Peter Gill, James Curran, and Keith Elliot. Supplementary material. Published online at Nucleic Acids Research, 2005.